Regular Article

ORGANIC CHEMISTRY RESEARCH
Published by the
Iranian Chemical Society
www.orgchemres.org

Org. Chem. Res., Vol. 7, No. 1, 61-76, March 2021.
DOI: 10.22036/org.chem.2022.297270.1261

An Investigation on the QSAR Modeling of Carfilzomib Derivatives Using Monte Carlo Method and Novel Modelling-optimization Approach

R. Sayyadi Kord Abadi*, O. Alizadeh and G. Ghasemi

Department of Chemistry and Chemical Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran

(Received 28 January 2021, Accepted 7 February 2022)

The activity of 25 different carfilzomib derivatives was estimated using multiple linear regression (MLR), artificial neural network (ANN), and genetic algorithm (GA) as well as simulated annealing (SA) and imperialist competitive algorithm (ICA) as optimization methods. When comparing the results of gas phase from MLR-MLR, MLR-GA, SA-ANN, and GA-ANN techniques, we observed root mean sum square error (RMSE) of 0.290, 0.0482, 0.0294, and 0.0098, respectively, for combinations of modelling-optimization methods. There was a high predictive ability for the MLR-ICA model with the best number of empires/imperialists (nEmp = 50) and nEmp = 100 with RMSE of 0.00996 in the gas phase. The MLR-ICA method revealed that RDF 075m, MATS1m, F04[N-O], O-059, F09[C-O], and Mor21p are the most important descriptors. From Monte Carlo simulations, it was found that the presence of double bonds, the absence of halogens, oxygen connected to a double bond, sp^2 carbon, and double bonds with ring, branching, and nitrogen are important molecular features affecting the biological activity of the drug. It was concluded that simultaneous utilization of MLR-ICA, GA-ANN, and Monte Carlo method can lead to a more comprehensive understanding of the relation between physicochemical, structural, or theoretical molecular descriptors of drugs to their biological activities and facilitate designing of new drug.

Keywords: Carfilzomib, Antitumor drugs, QSAR, Genetic Algorithm, Monte Carlo method

INTRODUCTION

Carfilzomib is a second-generation proteasome inhibitor approved by the US food and drug administration (FDA) for relapsed/refractory multiple myeloma (RR-MM). The drug is being investigated in various combinations in relapsed disease, newly diagnosed MM (ND-MM) as well as other hematological and non-hematological malignancies [1-3], but clinical benefit to date is limited to MM. Carfilzomib is characterized by a rapid onset of action and can be safely administered in patients with renal impairment, a frequently encountered situation in patients with MM. Overall, carfilzomib has broadened the treatment options for patients with progressive MM and its presentation was suggested to prolong patients' survival [4], although this needs to be validated in prospective studies [5,6].

*Corresponding authors. E-mail: Sayyadi@iaurasht.ac.ir

QSAR methods are mathematical equations establishing a relationship between chemical structures and biological activities [7-10]. There are several QSAR techniques that include multiple linear regression (MLR), simulated annealing (SA) [11-13], genetic algorithm (GA) [14], partial least squares (PLS); they can be used in the development of a quantitative relationship between the structural descriptors and the property [15,16]. Imperialist Competitive Algorithm (ICA) starts with an initial population called countries; it was proposed by Atashpaz-Gargari and Lucas [17] and is a new population-based optimization algorithm that has recently been introduced for dealing with different kinds of optimization problem [18-20,10].

CORAL has been proposed as a competent software for the QSAR studies. It uses Monte Carlo method to find the most important simplified molecular input-line entry system (SMILES)-based descriptors and calculate their correlation weights to predict an endpoint (*e.g.*, -log(IC_{50})). SMILES are

lines of symbols, representing the molecular structure [21,22]. In the present study, CORAL software, MLR-ICA approach, and various QSAR models including SA and genetic algorithm GA are utilized to select the best descriptors for the important prediction of inhibitory activity of carfilzomib derivatives. At last, the models are compared.

THEORY AND COMPUTATIONAL METHODS

Linear and Non-linear QSAR Study

Geometrical optimizations of carfilzomib derivatives were carried out with B3LYP/6-31G level of theory by Gaussian 03W [24,24] software. Computations based on density functional theory (DFT) are the main tools in contemporary computational chemistry. Although B3LYP functional underestimates the reaction barrier heights, yields too low bond dissociation enthalpies, and fails to bind van der Waals systems, it is very promising in a number of areas of chemistry, such as geometry, stability of complexes, molecular spectroscopic properties, electronic structures, and bonding characters [26-29]. For open-shell calculations UB3LYP is applied that performs the best of all the functionals in terms of geometry and prediction of the singlet–triplet energy gap. The 6-31G basis set is a standard, split-valence double-zeta basis set, to describe the core and valence orbitals [26].

3225 descriptors were calculated by the Dragon program [30,31] categorized in MoRSE [32,33], RDF [33], topological, geometrical, GETAWAY [16] auto-correlations [33], and WHIM [23] groups. The descriptors with the same values for at least 70% of the carfilzomib derivatives in the data set were removed and then the descriptors with a correlation coefficient of less than 0.25 with the dependent variable -log (IC50) (empirical negative logarithm of half maximal inhibitory concentration) for carfilzomib derivatives [34] were considered redundant and subsequently removed [35]. Following these two steps, the number of descriptors was reduced to 1198 descriptors in gas phase; as a result, a stepwise multiple linear regression procedure based on the forward-selection and backward-elimination techniques was used for the inclusion or the rejection of the descriptors in the screened method and 6 descriptors were selected as the best descriptors. Low standard deviation, high correlation coefficient (R), and root mean sum square errors (RMSE) [36] are characteristics of an ideal model, where the RMSE is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_o)^2}{n}} \quad (1)$$

where y_i is the desired output, y_a is the predicted value by method, and n is the number of molecules in our data set.

A stepwise multiple linear regression procedure based on the forward-selection and backward-elimination techniques was used for the rejection of descriptors in the linear models. The MLR model maps independent variables X to a dependent variable (response) Y using the following relation:

$$Y = W_1X_1 + W_2X_2 + \ldots + W_pX_p \quad (2)$$

where W_i is the coefficient of the regression. An ideal method is a method that has low standard deviation, high correlation coefficient (R^2), the minimum number of independent variables, high predictive power, and a high F-statistic value [37]. To establish the simulated annealing artificial neural network (ANN) (SA-ANN), MLR-GA, and GA-ANN models, 1198 descriptors in the gas phase were fed to the neural network algorithm to select the best descriptors so that 80%, 10%, and 10% of data sets in these models were randomly chosen as training, validation, and test sets, respectively. The networks were trained by using the TSET members with Levenberg-Marquardt algorithm [38]. Such networks were supposed to identify the non-liner relationship between the structural descriptors and inhibitory activity of carfilzomib derivatives. This study also used three neurons in the hidden layer of the ANN approaches (Fig. 1). All the calculations in our study were carried out in MATLAB environment (2014a, The Mathworks, Inc.).

The 1198 SPSS [39] screened descriptors were used as the feed to MLR-ICA approach as the population matrix in order to find the best descriptors for the gas phase. The numbers of the most effective descriptors (6 for the gas phase)

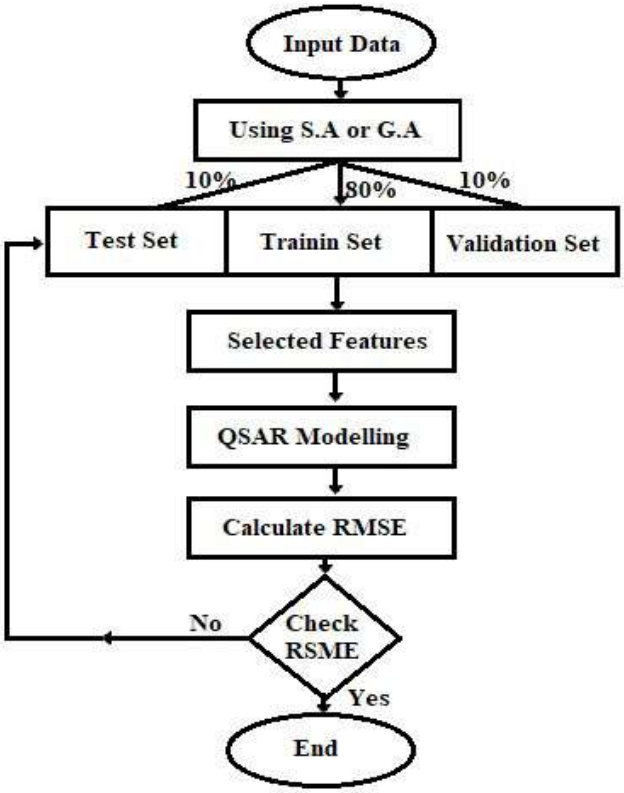

Fig. 1. The employed approach for finding optimum descriptors of the ANN methods.

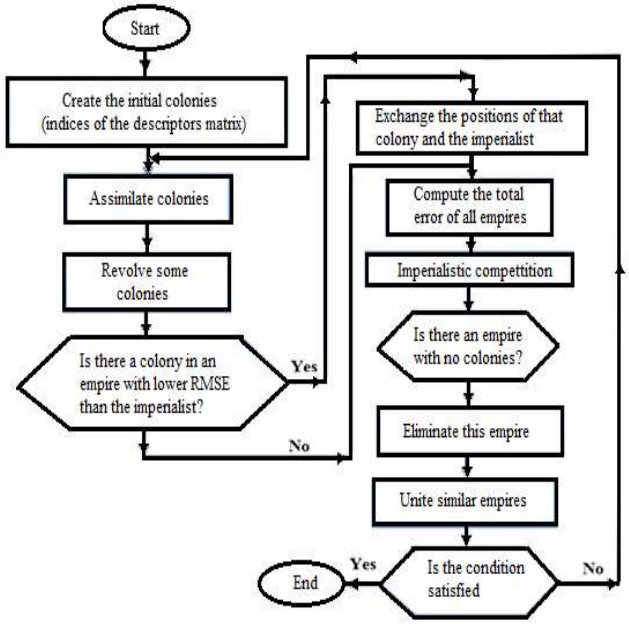

Fig. 2. Flowchart of the imperialist competitive algorithm.

were chosen by a stepwise multiple linear regression procedure in this work. The developed algorithm of this work is depicted in Fig. 2.

The procedure begins from random points (matrix indices of descriptors) called the initial countries that are the counterpart of chromosomes in GA and it is a set of values of a candidate solution for the optimization problem. The empires are sub-populations of the countries. Assimilation, which can be considered as a primitive form of particle swarm optimization [17,40,41] moves all non-best countries (called colonies) in an empire toward the best country (called imperialist) in the same empire to find the colonies with the lowest error (RMSE of MLR-predicted -log(IC50) *versus* the empirical values). Different number of empires (nEmp) were investigated to obtain the least RMSE and highest R^2. The number of decision variables (nDes) and number of empires/ imperialists (nEmp) were considered 6 and 10, 20, 30, respectively.

Monte Carlo Method

CORAL [42] software was used for calculation of descriptor correlation weight (DCW) of the 25 carfilzomib compounds with a hybrid optimization scheme including hydrogen-suppressed molecular graph (HSG), hydrogen-filled graphs (HFG), and SMILES representations of molecular structures. Modelling using CORAL software was carried out for thresholds of 1 up to 3 and 100 epochs (*i.e.*, an overall number of 900 runs were performed) [43]. The SMILESand Graph-based optimal descriptors are achieved using the following equations [44]:

$$DCW(T,Nepoch)^{SMILES} = \alpha\sum CW(Sk) + \beta\sum CW(SSk) + \gamma\sum CW(SSSk) + xCW(NOSP) + yCW(HALO) + zCW(BOND) \quad (3)$$

$$DCW(T,Nepoch)^{Graph} = \sum CWAk + \alpha\sum CW({}^{0}ECk) + \beta\sum CW({}^{1}ECk) + \gamma\sum CW({}^{2}ECk) + \delta\sum CW({}^{3}ECk) \quad (4)$$

where, *S*k, *SS*k, and *SSS*k denote component SMILES

attributes. The NOSP and HALO are presence or absence of chemical elements nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, and bromine, respectively. Double (=), triple (#), or stereo chemical bonds (@ or @@) are shown with the "BOND". *Ak* in Eq. (4) indicates the occurrence of the C, N, O atoms in the HSG and HFG molecular graphs. The α, β, γ, and δ coefficients and combinations of their values are used to define various versions of the graph-based optimal descriptors and can be 1 or 0. The hybrid objective function for finding the optimal descriptors is defined as:

$$\mathrm{DCW}\ (T,Nepoch)^{\mathrm{Hybrid}} = \mathrm{DCW}\ (T,Nepoch)^{\mathrm{SMILES}} + \mathrm{DCW}\ (T,Nepoch)^{\mathrm{Graph}} \quad (5)$$

RESULT AND DISCUSSION

Molecular Descriptors Generation with Linear and Nonlinear Methods

Twenty-five different carfilzomib derivatives were selected as a sample set, and the geometry of the compounds was optimized using Gaussian 09W at B3LYP/6-31G. The optimized structures of the studied derivatives are given in Fig. 3 and their structural parameters involving C-N and C=O bond lengths (in Å) as well as their energy values (in Hartree) are given in Table S1 of supplementary file.

The best selected descriptors utilizing MLR-MLR, SA-ANN, MLR-GA, and GA-ANN methods in the gas phase are shown in Table 1. These parameters relate the structure to the activity of the optimized carfilzomib derivatives.

In MLR-PCR, MLR-PLS1, and MLR-MLR models, the best descriptors were selected using MLR procedure of SPSS [39] software in three steps described in theory and computational methods section. Thereafter, the selected descriptors were employed as input in unscramble (V.9.7) software and statistical parameters were calculated using PCR, PLS1, and MLR models (Table 2).

When the MLR-MLR model was utilized, the RMSE of the predicted activity was found to be 0.2081 in the gas phase. In addition, the correlation coefficient (R^2) calculated for the PSET was 0.8850 in the gas phase. It was demonstrated that MLR-MLR method is more effectively than other linear methods (MLR-PLS1 and MLR-PCR); see Table 2 for data.

To establish the SA-ANN, MLR-GA, and GA-ANN methods, the 1198 descriptors in the gas phase were fed to the NN to select the best descriptors as described in theory and computational methods section. The statistical parameters of all non-linear QSAR models are shown in Table 3.

Table 1. The Best Selected Descriptors Using QSAR Methods

MLR-MLR	MLR-GA	SA-ANN	GA-ANN
RDF145u	EPs1	RDF045v	RDF120u
Mor17v	RDF015p	Tp	piPC04
R5u+	MATs6e	MATs2e	Mor27u
R5v	EEig08r	RDF040m	O-059
R7e	Mor09m	Mor28m	ATS6v
FO1[C-N]	H6p	RDF125m	Mor26p

Table 2. Statistical Parameters of Different Linear QSAR Models in Gas Phase

QSAR Model	R^2	RMSE
MLR-PLS1	0.8314	0.2520
MLR-PCR	0.8248	0.2569
MLR-MLR	0.8850	0.2081

Table 3. Statistical Parameters of Different QSAR Models in Gas Phase

QSAR Model	Predicted		Train	
	R^2	RMSE	R^2	RMSE
SA-ANN	0.9276	0.0294	0.9477	0.0269
MLR-GA	0.874	0.0482	0.8561	0.0594
GA-ANN	0.9746	0.0098	0.9742	0.0088

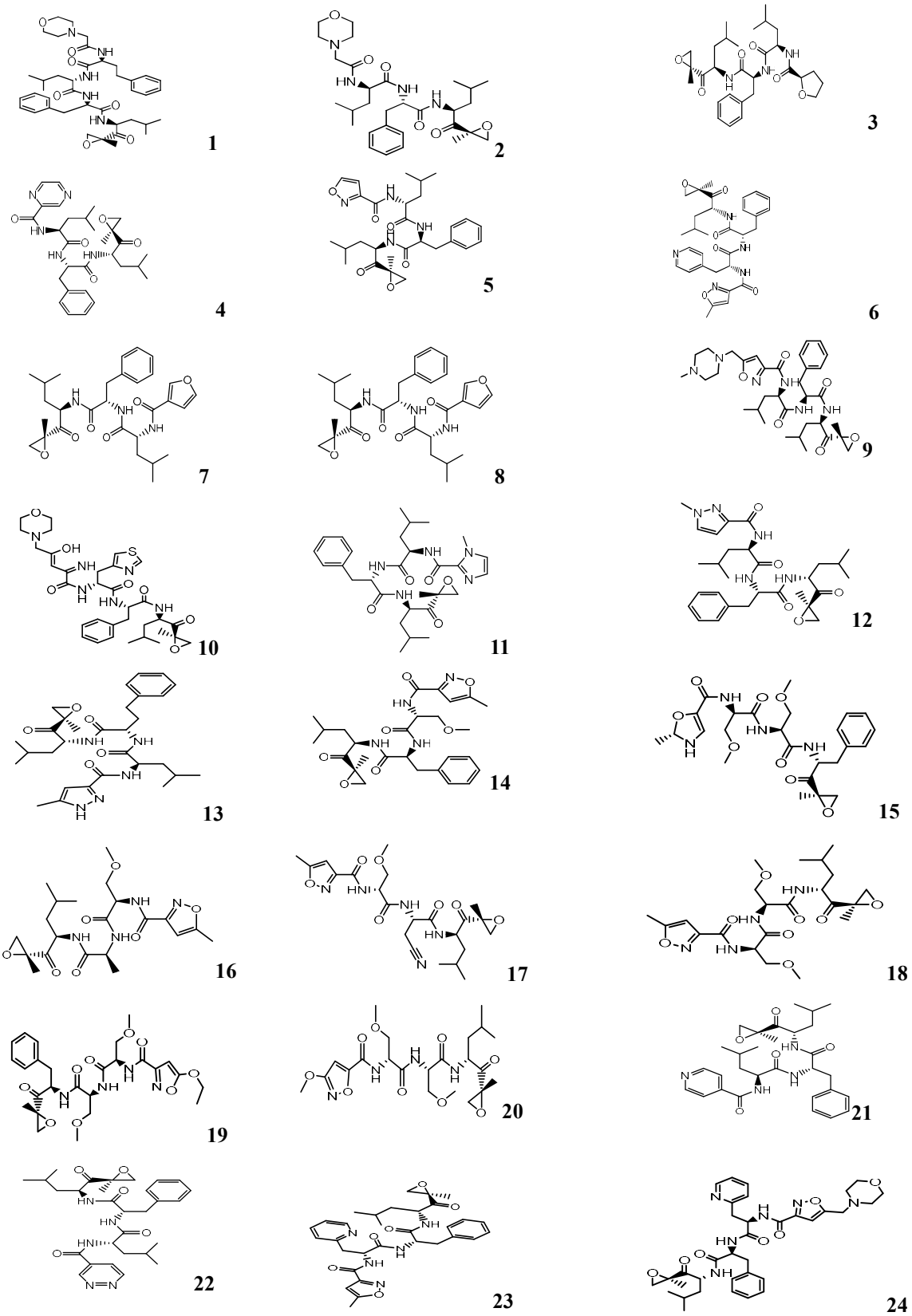

Fig. 3. Schematic figure of optimized structures of target compounds used to build QSAR models with B3LYP/6-31G.

Table 4. The Best Selected Descriptors Using GA-ANN Method in Gas Phase

Description	Definition	Type
RDF120u	Radial Distribution Function-120/unweighted	RDF descriptors
piPC04	Molecular multiple path count of order 04	Walk and Counts descriptors
Mor27u	3D-MoRSE-Signal 27/unweighted	3D-MoRSE descriptors
O-059	AI-O-AI.	Atom-center Fragments descriptors
ATS6v	Broto-Moreau autocorrelation of a topological structure -lag 6/Weighted by atomic van der Waals volume	2D autocorrelations descriptors
Mor26p	3D-MoRSE-signal26/Weighted by atomic polarizabilities	3D-MoRSE descriptors

In the GA-ANN method, the RMSE and R^2 values were calculated as 0.0098 and 0.9746, respectively, in the gas phase for the predicted activity of carfilzomib derivatives. Therefore, GA-ANN method is more accurate than the other models and as such, only the descriptors utilized in this method were evaluated in this study as shown in Table 4.

In Table 4, ATS6v is 2D-autocorrelation that these descriptors describe the considered properties distributed along the topological structure [33]. Topological index mathematically encodes information regarding the structure of molecules, which has been depicted as graphs and are often sensitive to size, shape, branching, cyclicity, and to a certain extent, the electronic characteristics of molecules [33]. RDF120u descriptors are independent of the number of atoms, *i.e.*, the size of a molecule, and provide further valuable information, for instance, about bond distances, ring types planar and non-plane systems, and atom types [33]. piPC04 is walk and path descriptor. The molecular multiple path counts (piPCk) are defined as path counts weighted by the bond order. It should be noted that aromatic bonds are assigned a bond order of 1.5. In order to match Dragon, a similar logarithmic transformation used again to calculate piPCk [33]. The Mor27u and Mor26p descriptors with their fixed length representation of 3D molecular structure allow a comparison of datasets of compounds of different sizes, with different number of atoms. The presence of a MoRSE descriptor indicates that the size of the inhibitor compound has certain effect on the extent of the interaction between the enzyme and compound; O-059 is atom-center fragment descriptor that represents the local atomic environments [33].

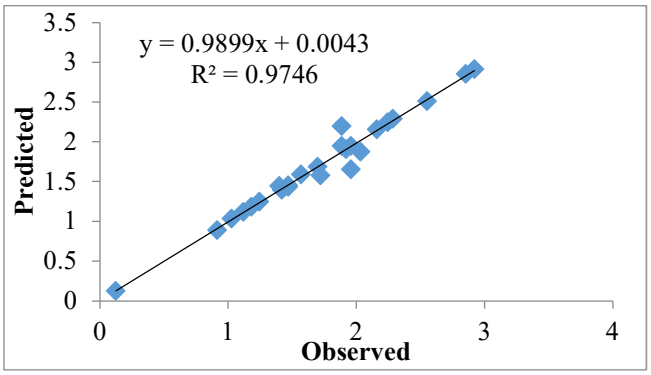

Fig. 4. Plot between observed and predicted -log(IC50) using GA-ANN model in gas phase.

The plot showing the variation of observed -log(IC_{50}) *versus* predicted empirical negative logarithm of half maximal inhibitory concentration values in GA-ANN model are illustrated in Fig. 4. The figure implies that the developed model possesses a high correlation coefficient which indicates the experimental and predicted values are well correlated.

Considerable RMSE values result from the possible errors in experimental data employed in this study. As a result, RMSE highly depends on the range of the dependent variable [45]. As Table 5 lists, the values of -log(IC50) in our dataset are in the range of 0.124 to 2.94. The RMSE values of the predicted set in GA-ANN model were 0.0098 and 0.0202 in

Table 5. Observed and Predicted Values of -log(IC50) by Using GA

Compound	-logIC_{50} Observed	-logIC_{50} Calculated	Compound	-logIC_{50} Observed	-logIC_{50} Calculated
1	2.244	2.2443	14	2.032	1.876
2	1.244	1.248	15	1.181	1.183
3	1.469	1.454	16	1.721	1.576
4	0.124	0.125	17	1.398	1.446
5	1.569	1.589	18	1.959	1.948
6	2.553	2.510	19	2.854	2.853
7	1.469	1.434	20	1.959	1.654
8	1.886	1.946	21	1.699	1.687
9	1.027	1.033	22	1.921	1.909
10	2.161	2.159	23	1.886	2.197
11	0.914	0.888	24	2.284	2.290
12	1.119	1.1197	25	2.921	2.913
13	1.42	1.395			

the gas phase, respectively, which are acceptable in comparison with previous works [36].

Graphs predicted -log(IC50) of RDF120u, piPC04, Mor27u, O-059, ATs6v, and Mor26p descriptors in the gas phase *versus* the empirical negative logarithm of half maximal inhibitory concentration are demonstrated in Fig. 5 computed by using the MATLAB program. The charts in gas phase show that the empirical negative logarithm of half maximal inhibitory concentration value increases with increasing RDF120u, piPC04, Mor27u, O-059, ATs6v, and Mor26p structural and physicochemical descriptors; consequently,

-log(IC50) value reduces that confirms the aforementioned descriptors are the best descriptors in the gas phase. In other words, ring type planar and non-plane systems, aromatic bonds, size, number of atoms, van der Waals volumes, and dipole moment values should be maximized in designing new drugs. Approximately, 4.2 responses do not change in ATs6v (Weighted by atomic van der Waals volume) descriptor. However, the values from 4.2 to 4.4 lead to an increased empirical negative logarithm of half maximal inhibitory concentration rate which is seen in the bar chart.

Molecular Descriptors Generation with MLR-ICA Approach

As a first trial, 1000 number of iterations were done to find the most powerful empires, i.e., the best descriptors. A plot of the best cost values versus the number of iterations is represented in Fig. 6. It implies that there is no variation in the best cost (MSE) after about 300 iterations. However, in order to ensure that the best descriptors are captured, the number of iterations for the rest of computations was set to 500.

The effects of the number of selected descriptors on the chosen descriptors and the prediction quality (according to R^2 and RMSE) was investigated and the results are shown in Table 6. As it is expected, the model's accuracy regarding to R^2 and RMS increases by increasing the number of model parameters (descriptors in this case).

In order to choose the most suitable number of empires, the model was run using different number of empires and the results are demonstrated in Table 7. According to Tables 6 and 7, the optimum number of empires was chosen as 50 with 6 number of the best descriptors. The chosen descriptors using MLR-ICA approach are presented in Table 8.

F04 [N-O] and F09 [C-O] (Table 8) are topological descriptors that are often sensitive to the electronic characteristics of molecules and size, shape, branching, and 2D-frequency fingerprints (fragment descriptors are representations of local atomic environments) descriptors,

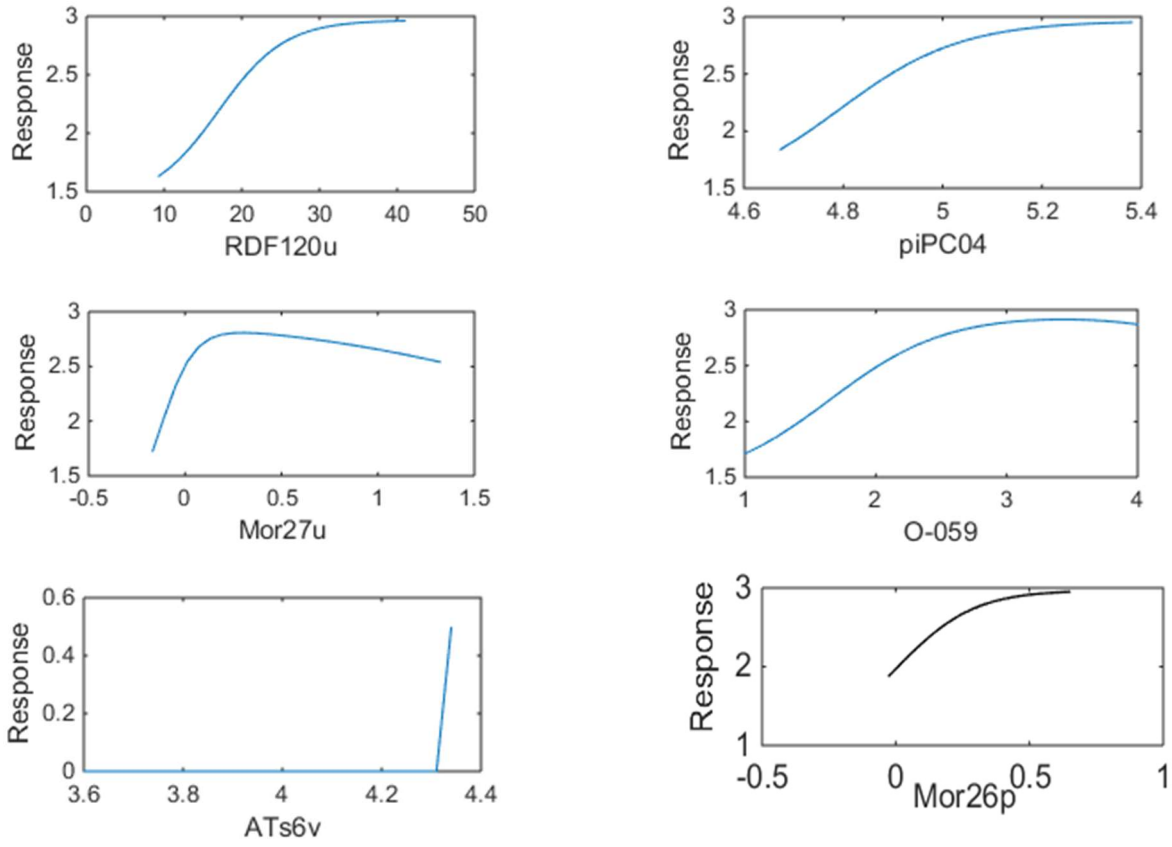

Fig. 5. Plot between -logIC50 experimental (Response) *versus* of the RDF120u, piPC04, Mor27u, O-059, Ats6v, and Mor26p descriptors in GA-ANN method.

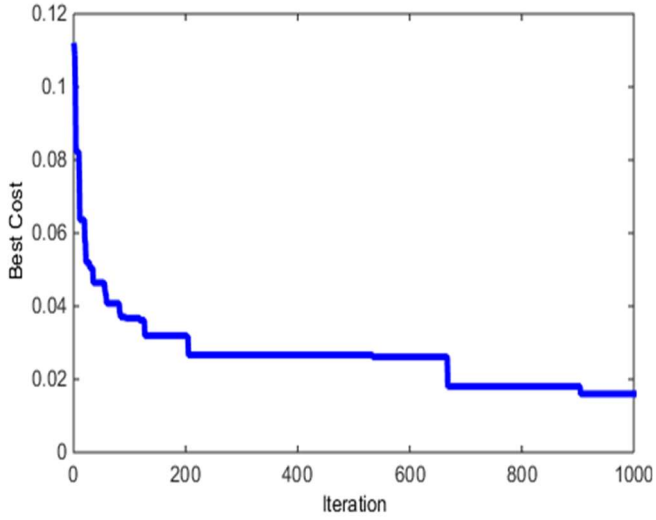

Fig. 6. Plot between best cost values *versus* the variation of iteration with nVar = 6, nEmp = 30 in gas phase.

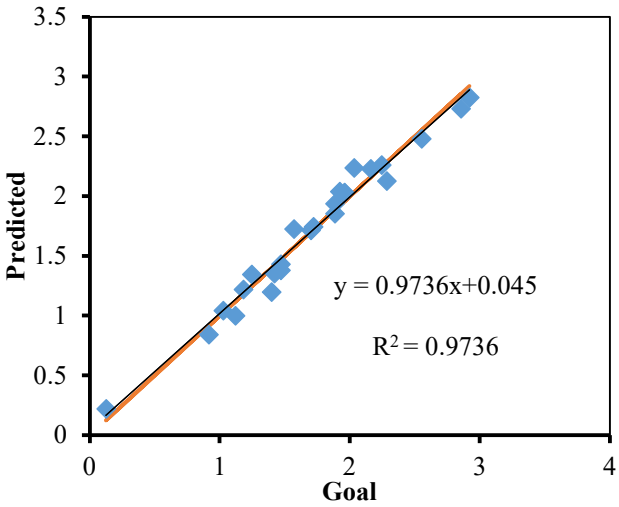

Fig. 7. Plot between predicted values *versus* goal with nVar = 6 and nEmp = 50, in gas phase.

Table 6. Statistical Parameters of ICA-MLR Models in Gas Phase (Max. It = 5000)

nVar-nEmp	R^2	RMSE	nVar-nEmp	R^2	RMSE
1-10	0.5554	0.1674	4-20	0.887	0.0425
2-10	0.7540	0.0843	5-20	0.9079	0.0347
3-10	0.826	0.0655	6-20	0.9597	0.0152
4-10	0.8706	0.0487	1-30	0.5554	0.1674
5-10	0.9368	0.0239	2-30	0.7540	0.0843
6-10	0.921	0.0297	3-30	0.826	0.0655
1-20	0.5554	0.1674	4-30	0.8682	0.0496
2-20	0.7540	0.0843	5-30	0.9079	0.0347
3-20	0.8187	0.0183	6-30	0.9694	0.0115

Table 7. Statistical Parameters of ICA-MLR Models in Gas Phase with Different nEmp (Max.It = 1000)

nVar-nEmp	Predicted (Gas phase)	
	R^2	RMSE
6-40	0.9520	0.0181
6-50	0.9736	0.009916
6-60	0.9587	0.01551
6-70	0.9452	0.02061
6-80	0.9623	0.01421
6-90	0.9626	0.0141
6-100	0.9552	0.0169

Table 8. The Best Selected Descriptors Using MLR-ICA Method with nDes = 6 and nEmp = 50 in Gas Phase

Descriptor	Definition	Type
RDF075m	Radial Distribution Function-7.5/Weighted by atomic masses	RDF descriptors
MATS1m	2D autocorrelation-lag 1/weighted by atomic masses	2D autocorrelation
F04[N-O]	Frequency of N-O at topological distance 04	2D Frequency fingerprints
O-059	Al-O-Al	Atom Centered Fragments
F09[C-O]	Frequency of C-O at topological distance 09	2D Frequency fingerprints
Mor21p	3D-MoRSE-signal 21/weighted by atomic polarizabilities	3D-MoRSE descriptors

Plots of the RDF 075m, MATS1m, F04[N-O], O-059, F09[C-O], and Mor21p descriptors *versus* empirical negative logarithm of half maximal inhibitory concentration (-logIC50, respectively [33]. A plot of the predicted *versus* empirical values of -log(IC50) is depicted. The figure implies that the developed model possesses a high correlation coefficient indicating well correlation of the experimental and predicted values.

Response) were plotted using MATLAB program and are displayed in Fig. 8. Charts show that an increase in the amount of the F04[N-O] descriptor brought about an increase in the amount of response (-logIC50) to 3 and then the amount of response (-logIC50) decreases. Increase in RDF075m, MATS1m, O-059, and Mor21p descriptors results in an increase in response. As the MATs8v (size, shape, weighted by atomic masses, atom centered fragments, and dipole moment should be maximized in designing new drugs.

Results of the Monte Carlo Method

The statistical parameters of the models obtained using molecular graphs (HSG) and SMILES are shown in Table 9. branching, and cyclicity) descriptor increased to near 22, no change in response was observed. Thus, during this period, a bar can be seen in the response. Therefore ring types planar and non-plane systems and atom types, branching in Performance of the models were compared with each other by the criterion of the predictability in test set (Rm^2) which should be larger than 0.5 [46], correlation coefficient

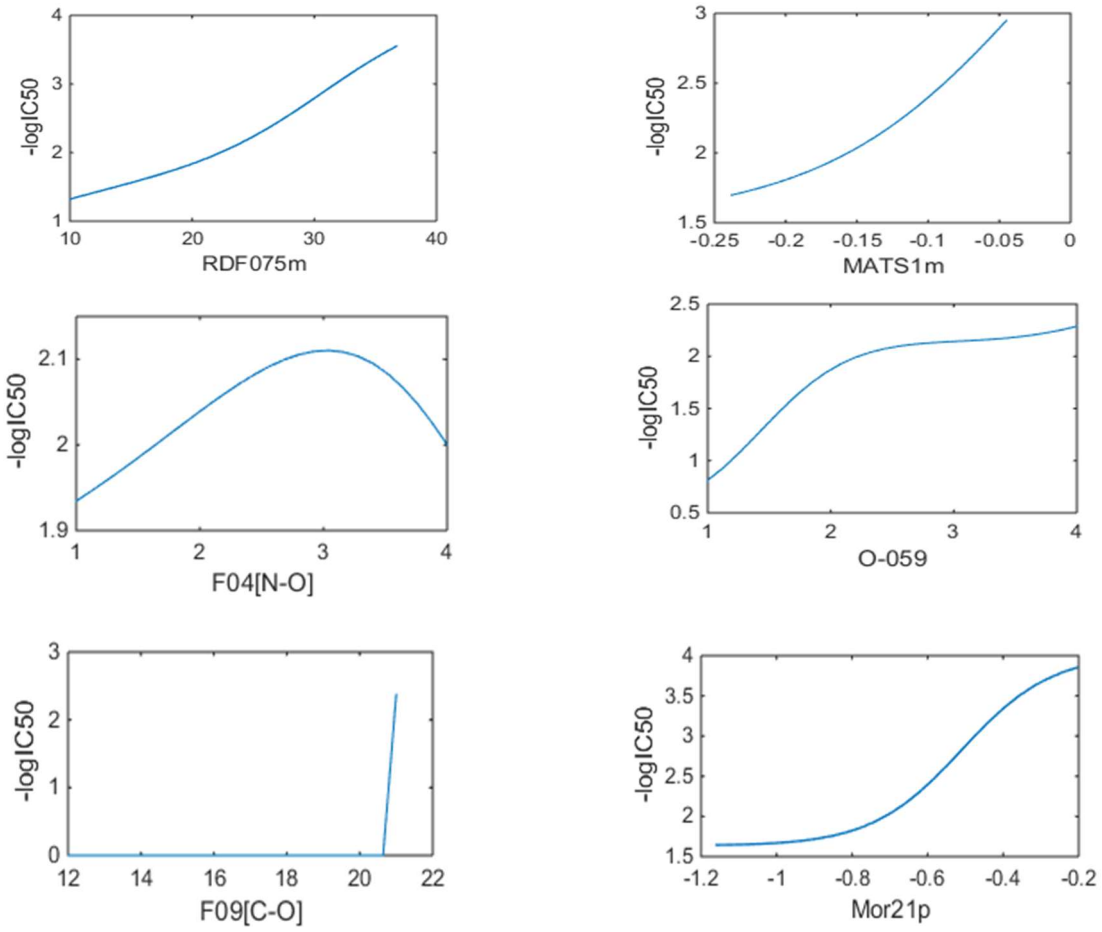

Fig. 8. Plot between -logIC50 experimental (Response) *versus* of the RDF 075m, MATS1m, F04[N-O], O-059, F09[C-O], and Mor21p descriptors in the gas phase.

(R^2) in each set, cross-validated correlation coefficient (Q^2), and standard error of estimation (s). The difference between R^2m and R'^2m values (ΔRmTEST) was used as another criterion in this issue. The results with threshold of 3 and probe 3 that are the best ones in comparison to the other amounts are presented in Table 9.

The variation of correlation coefficient (test set) with respect to threshold and the number of epochs is plotted in Fig. 9 that confirms 3 and 60 are the most appropriate values for threshold and number of epochs, respectively.

Table 9. The Split Models in Monte Carlo Method

Split 1: (T = 3, prob = 3)
-logIC50 = -11.1566584 (± 0.4334670) + 0.0734383 (± 0.0025502) * DCW (3,100)
n = 10, R^2 = 0.8196, Q^2 = 0.7591, s = 0.706 (training set)
n = 8, R^2 = 0.9998, Q^2 = 0.9995, s = 0.870 (calibration set)
n = 7, R^2 = 0.9254, Q^2 = 0.7216, s = 0.84 (test set), R^2m TEST = 0.9244

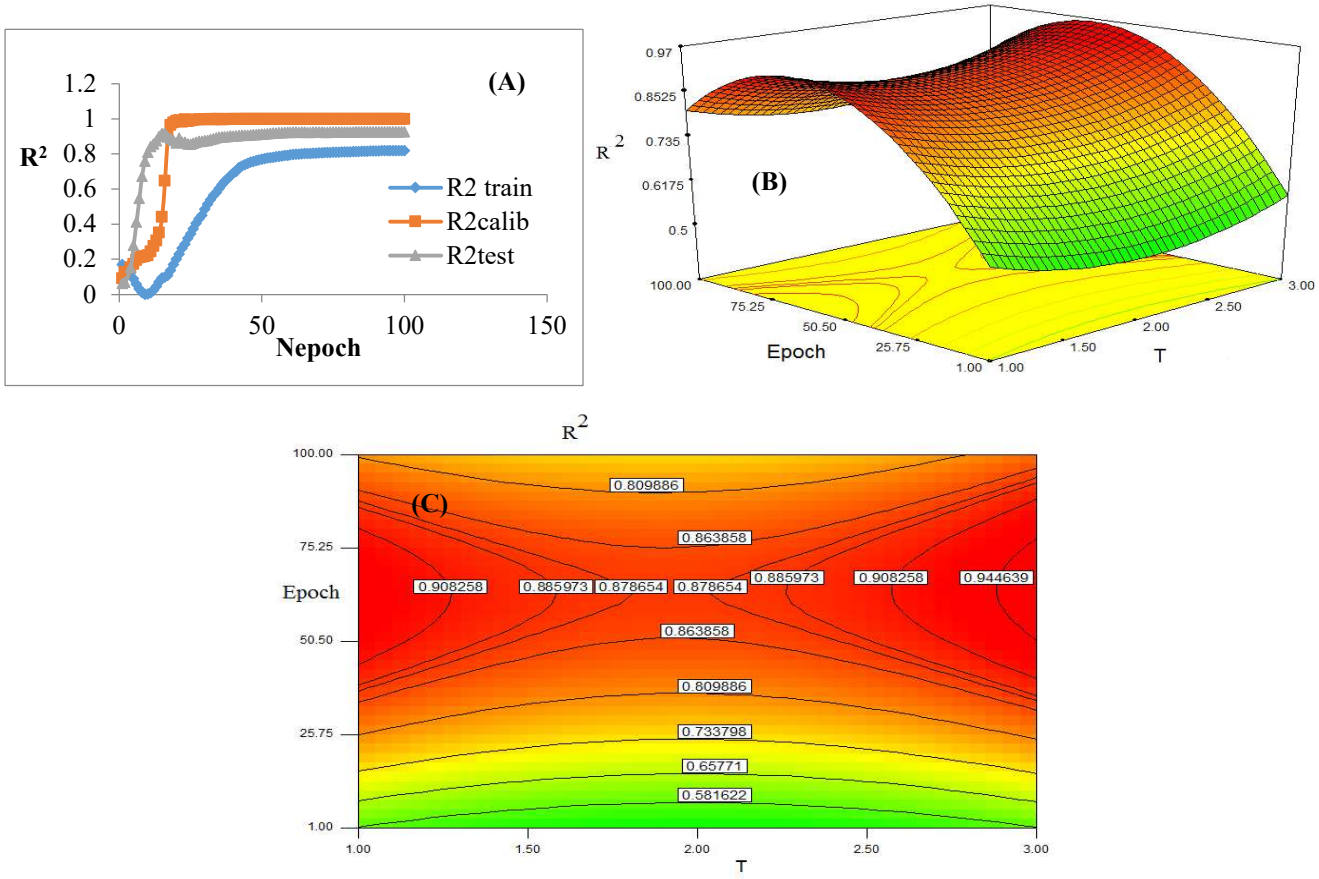

Fig. 9. The variation of correlation coefficient for test set by threshold and number of epochs. (A) Effects of the number of epochs. (B) 3-D surface plot of R^2 according to the threshold and the number of epochs. (C) Contour plots of R^2 according to the threshold and the number of epochs.

The distribution of SMILES notations in the train, calibration, and test sets are reported in Table 10. The experimental and calculated activities (-logIC50) for the sequence of compounds are plotted against each other in Fig. 10. A strong correlation between the calculated and empirical values of -logIC50 can be observed that approves the appropriateness of the developed model.

Table 10. SMILES Notations of 25 Compounds of carfilzomib, their Set of Train, Calibration, and Test

Compound	SMILES	Set
1	CC(C)CC(N:C(=O)C(CCC:1:C:C:C:C:C:1)N:C(=O)CN2CCOCC2)C(=O):NC(CC:3:C:C:C:C:C:3)C(=O):NC(CC(C)C)C(=O)C4(C)CO4~	Train
3	CC(C)CC(NC(=O)C1CCCO1)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Train
6	CC(C)CC(NC(=O)C(CC1=CC=CC=C1)NC(=O)C(CC2=CC=NC=C2)NC(=O)C3=C[C-](C)ON3)C(=O)C4(C)CO4	Train
8	CC(C)CC(NC(=O)C1=NOC(=C1)C[N]2C=NC=N2)C(=O)NC(CC3=CC=CC=C3)C(=O)NC(CC(C)C)C(=O)C4(C)CO4	Train
9	CC(C)CC(N:C(=O)C:1:NOC(=C:1)CN2CCN(C)CC2)C(=O):NC(CC:3:C:C:C:C:C:3)C(=O):NC(CC(C)C)C(=O)C4(C)CO4	Train
13	CC(C)CC(NC(=O)[C-]1NNC(=C1)C)C(=O)NC(CCC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Train
16	COCC(N:C(=O)C:1:NOC(=C:1)C)C(=O):NC(C)C(=O):NC(CC(C)C)C(=O)C2(C)CO2	Train
20	COCC(NC(=O)C(COC)NC(=O)C(/[O-])=C/C(=N)OC)C(=O)NC(CC(C)C)C(=O)C1(C)CO1	Train
23	CC(C)CC(NC(=O)C(CC1=CC=CC=C1)NC(=O)C(CC2=CC=CC=N2)NC(=O)[C-]3NOC(=C3)C)C(=O)C4(C)CO4	Train
24	CC(C)CC(NC(=O)C(CC1=CC=CC=C1)NC(=O)C(CC2=CC=CC=N2)NC(=O)C(=[N-])[CH-]C(=O)CN3CCOCC3)C(=O)C4(C)CO4	Train
4	CC(C)CC(N:C(=O)C:1:C:N:C:C:N:1)C(=O):NC(CC:2:C:C:C:C:C:2)C(=O):NC(CC(C)C)C(=O)C3(C)CO3	Calib
5	CC(C)CC(NC(=O)C1=NOC=C1)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Calib
7	CC(C)CC(NC(=O)[C-]1NNC(=C1)C)C(=O)NC(CCC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Calib
10	CC(C)CC(NC(=O)C(CC1=CC=CC=C1)NC(=O)C(CC2=CSCN2)NC(=O)C(/N)=C/C(=O)CN3CCOCC3)C(=O)C4(C)CO4	Calib
11	CC(C)CC(NC(=O)[C-]1NC=CN1C)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Calib
12	CC(C)CC(NC(=O)[C-]1NN(C)C=C1)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Calib
15	COCC(NC(=O)C(COC)NC(=O)C1=CNC(C)O1)C(=O)NC(CC2=CC=CC=C2)C(=O)C3(C)CO3	Calib
18	COCC(NC(=O)C(COC)NC(=O)[C-]1NOC(=C1)C)C(=O)NC(CC(C)C)C(=O)C2(C)CO2	Calib

Table 10. Continued

2	CC(C)CC(N:C(=O)CN1CCOCC1)C(=O)NC(CC:2:C:C:C:C:C:2)C(=O):NC(CC(C)C)C(=O)C3(C)CO3	Test
14	COCC(NC(=O)[C-]1NOC(=C1)C)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Test
17	COCC(NC(=O)[C-]1NOC(=C1)C)C(=O)NC(CC#N)C(=O)NC(CC(C)C)C(=O)C2(C)CO2	Test
19	CCO[C-]1ONC(=C1)C(=O)NC(COC)C(=O)NC(COC)C(=O)NC(CC2=CC=CC=C2)C(=O)C3(C)CO3	Test
21	CC(C)CC(NC(=O)C1=CC=NC=C1)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Test
22	CC(C)CC(NC(=O)C1=CC=NN=C1)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Test
25	COCC1=C[C-](NO1)C(=O)NC(CCC(C)C)C(=O)NC(CC2=CC=CC=C2)C(=O)NC(CC(C)C)C(=O)C3(C)CO3	Test

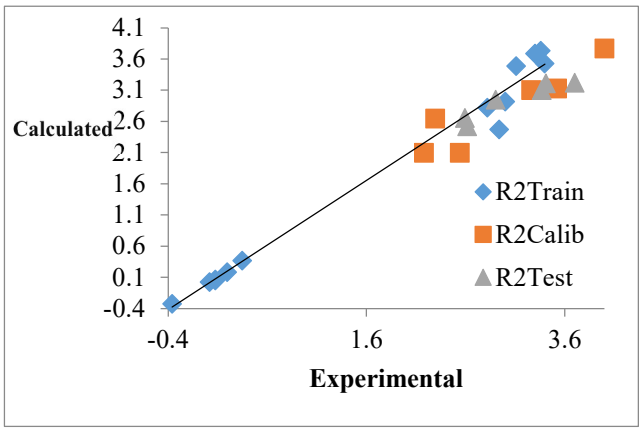

Fig. 10. Correlation between experimental and predicted -logIC50 calculated using Eq. (4).

Molecular features are sorted according to their correlation weights and are given in Table 11. Molecular features with negative correlation weights are omitted due to their inverse effect on the -logIC50 value. The higher the correlation weigh of a molecular feature, the lower the value of IC50; therefore, the feature is more significant. Definitions of the molecular features are given in Table 12. According to Table 11, presence of cyclic rings, presence of double bond connected to branching, absence of halogens, presence of oxygen connected to double bond, presence of sp^2 carbon connected to a double bond, presence of double bond with ring, presence of branching, and presence of nitrogen are the important molecular features that might be considered in designing new drugs.

CONCLUSION

The nonlinear feature selection methods were shown to be better than their linear methods, and the results of GA-ANN model were more precise than the other applied non-linear models. The results also confirm that the empirical negative logarithm of half maximal inhibitory concentration (-logIC50) value increases with increasing RDF120u, piPC04, Mor27u, O-059, ATs6v, and Mor26p descriptors in GA-ANN method and RDF075m, MATS1m, O-059, and Mor21p descriptors in ICA-MLR method. Therefore, the half maximal inhibitory concentration (IC50) value reduces. In Monte Carlo method, the structural descriptors are important. Precise setting of these physicochemical and structural descriptors can reduce IC50. The obtained results can be employed for designing new anti-cancer drugs. Additionally,

Table 11. SMILES Attributes with Positive Correlation Weights for Split 1

SMILES attributes	CWs	SMILES attributes	CWs
1.....	6.40015	NNC-O...110	5.60209
=...(.....	4.00484	C...=.....	4.38454
:BOND10000000	5.06631:	C5....H.1...	4.77323
EC0-C...4...	5.12481	=...2.....	8.14965
EC0-O...1...	6.23742	...(.....	4.82514
HALO00000000	6.68822	...1.....	4.42784
O...=.....	4.40119	N...:.....	4.42362
NNC-C...431.	6.17610		

Table 12. Definition of the Promoter of A_k

Attribute Ak	Comment
HALO00000000	Absence of F, Cl, Br
C...C.....	Presence of carbon-carbon bonds (sp3)
C…(…C…	SP3 Carbon atoms with branching
++++O---B2==	Presence of oxygen and double bonds
C…=…….	SP2 Carbon atom
(.....	Branching in molecular skeleton
O.....	Presence of oxygen
1.....	Presence of rings
++++N---B2==	Presence of nitrogen and double bond
=	Double bond
@	Stereo specific bond
#	Triplet bond

Monte Carlo method can be employed to find out the quality of the effects of structural descriptors on the biological activity of the studied drugs. It can be said that simultaneous use of Monte Carlo, GA-ANN, and ICA-MLR methods gives deeper and more comprehensive knowledge of the effect of molecular and structural descriptors on the activity of drugs and provides better insights for designing new drugs.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support provided by the Islamic Azad University of Rasht.

REFERENCES

[1] L.D. Wartman, M.A. Fiala, T. Fletcher, *et al.*, Leuk. Lymphoma 57 (2015) 1.

[2] S.P. Treon, C.K. Tripsas, K. Meid, *et al*., Blood 124 (2014) 503.

[3] K.P. Papadopoulos, H.A. Burris, M. Gordon, *et al.*, Cancer Chemother. Pharmacol. 72 (2013) 861.

[4] T.F. Wang, R. Ahluwalia, M.A. Fiala, *et al.*, Leuk. Lymphoma 55 (2014) 337.

[5] A.K. Stewart, S.V. Rajkumar, M.A. Dimopoulos, *et al.*, N Engl. J. Med. 372 (2015) 142.

[6] L. Vincenz, R. Jager, M. O'Dwyer, A. Samali, Mol. Cancer Ther. 12 (2013) 831.

[7] R. Sayyadi Kord Abadi, A. Alizadehdakhel, Rev. Roum. Chim. 63(2018) 931.

[8] R. Sayyadi Kord Abadi, A. Alizadehdakhel, Rev. Roum. Chim. 63 (2018) 171.

[9] R. Sayyadi Kord Abadi, A. Alizadehdakhel, S. Tajadodi Paskiabei, J. Korean Chem. Soc. 60 (2016) 225.

[10] R. Sayyadi Kord Abadi, A. Alizadehdakhel, S. Dorani Shiraz, S. Russ. J. Phy. Chem. B 11 (2017) 307.

[11] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671.

[12] V.O. Černý, J. Optimiz. Theory. App. 45 (1985) 41.

[13] L.M. Schmitt, Theor. Comput. Sci. 259 (2001) 1.

[14] D. Bertsimas, J. Tsitsiklis, Stat. Sci. 8 (1983) 10.

[15] R. Guha, J.R. Serra, P.C. Jurs, J. Mol. Graph. Model. 23

(2004) 1.

[16] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, P. J. Chem. Inf. Comput. Sci. 42 (2002) 693.

[17] Atashpaz-Gargari, E. Lucas, C. Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. in *IEEE* Congress on Evolutionary Computation. Singapore. (2007) 4661.

[18] S. Hosseini, A. Al Khaled, A. App. Soft Comput. 24 (2014) 1078.

[19] E. Shokrollahpour, M. Zandieh, B. Dorri, Int. J. Prod. Res. 49 (2011) 3087.

[20] M. Bagher, M. Zandieh, H. Farsijani, Int. J. Adv. Manuf. Tech. 54 (2011) 271.

[21] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, J. Comput. Chem. 32 (2011) 2727.

[22] A.A. Toropov, A.P. Toropova, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, J. Chemom. Intell. Lab. 109 (2011) 94.

[23] P. Gramatica, V. Consonni, R. Todeschini, Chemosphere 41 (2000) 763.

[24] E.B. DeMelo, M.M. Ferreira, Eur. J. Med. Chem. 44 (2009) 3577.

[25] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R.C. Fukuda, R. J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V.C. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V. G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09 (Gaussian, Inc., Wallingford CT, 2009). https://gaussian.com/glossary/g09/

[26] S. Baskaran, M.M Krishnan, R. Kumar, J. Molecul. Struct. 1224 (2021) 129.

[27] A. Abkari, I. Chaabane, K. Guidara, E. Physica, Low-dimensional Systems and Nanostructures 81 (2016) 136.

[28] I. Yılmaz, N. Acar-Selçuki, A. Şengül, J. Molecul. Struct. 1223 (2021) 129271.

[29] R.E. Hag, M.M Abdusalam, C. Aclian, H. Kayi, S. Özalp-Yaman, Polyhedron 170 (2019) 25.

[30] R. Todeschini, Milano Chemometrics, QSAR Group, http://www.disat.unimib.it/chem (2018).

[31] Dragon 3.0 Evaluation Version. Available online: http://www.disat.unimib.it/chm (accessed on 6 November 2008).

[32] J.H. Schuur, P. Selzer, J. Gasteiger, J. Chem. Inf. Comput. Sci. 36 (1996) 334.

[33] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, 2000.

[34] https://pubchem.ncbi.nlm.nih.gov

[35] M.H. Fatemi, S. Gharaghani, Bioorg. Med. Chem. 15 (2007) 7746.

[36] M. Nirouei, G. Ghasemi, P. Abdolmaleki, A. Tava.koli, S. Shariati, Indian J. Biochem. Biophys. 49 (2012) 202.

[37] M. Jalali-Heravi, M.F. Parastar, J. Chem. Inf. Comput. Sci. 40 (2000) 147.

[38] K. Levenberg, Q. App. Math. 2 (1944) 164.

[39]SPSS, Version 19, available at http://www.spssscience.com, 2010.

[40] E.A. Gargari, F. Hashemzadeh, R. Rajabioun, C. Lucas, J. Intelligent Compu. Cybernetics 1 (2008) 337.

[41] J.-L. Lin, Y.-H. Tsai, C.-Y. Yu, M.-S. Li, Algorithms 5 (2012) 433.

[42] http://www.insilico.eu/coral.

[43] J. Veselinović, A. Veselinović, A. Toropov, A. Toropova, I. Damnjanović, G. Nikolić, Scientific Journal of the Faculty of Medicine in Niš 31 (2014) 95.

[44] A. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikolić, Eur. J. Pharm. Sci. 48 (2013) 532.

[45] S.H. Sadat Hayatshahi, P. Abdolmaleki, M. Ghiasi, S. Safarian, FEBS Lett. 581 (2007) 506.

[46] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269.